The First Report of PRML

Liu Jitao
22371262@buaa.edu.cn

In this report, I am going to use 3 liner models and 2 non-linear models to fit a set of data consisting of 100 elements. Each part consists of the type of the models and the results are shown in figures. All codes are published on `https://github.com/KevinTJL/PRML-2025/tree/main/work_1`.

I. Results to Question1

A. LSM

LSM is widely used to fit one-dimensional functions. By using numpy library function we can easily build LSM model. Here is my result by using it to fit the training data and the testing data.

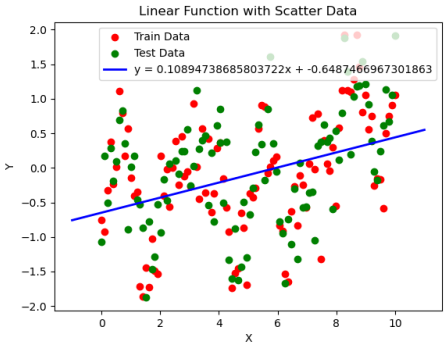

Fig. 1 This picture shows the fitness of LSM method with training data and test data

It is clear that the LSM method cannot fit non-linear data appropriately.

B. Gradient Descent

GM is another popular method in function fitting. In this method, I use the Euclidean norm as the loss function which can accurately reflect the error. The learning rate and the number of iterations are, respectively, 0.01 and 1000. The results are shown below. The polynomial to the highest power is 1.

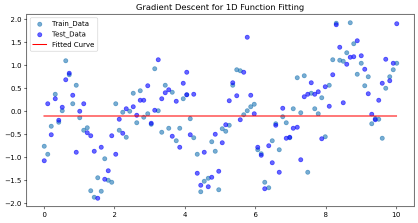

Fig. 2 This picture shows the fitness of GM method with training data and test data

It is clear that the GM method cannot fit non-linear data appropriately as well. Besides, if the polynomial to the highest power is greater than 1, there will not be any result due to gradient explosion.

C. Newton method

Newton method is much more often to fit high-dimension functions. Its performance in fitting one-dimension function is not good enough. Here is the result.

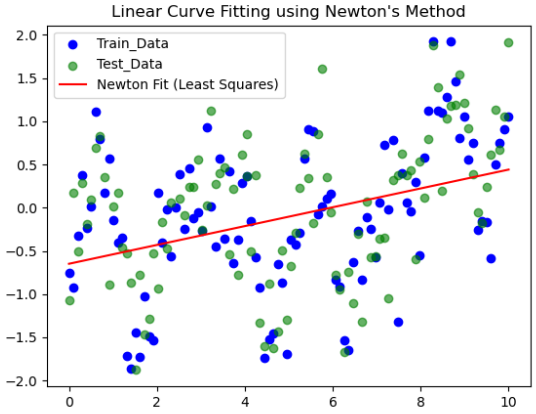

Fig. 3 This picture shows the fitness of GM method with training data and test data

The result is not good.

II. Results to Question2

As the previous results show, it is almost impossible to fit the given data with liner models. And I am going to use 2 non-liner models to solve the problem.

A. Random Forest Regression

Random Forest Regression (RF) is an ensemble learning method based on Decision Trees, which uses multiple decision trees for regression prediction and takes the average of multiple trees to improve the stability and generalization ability of the model. By using the RandomForestRegressor function from sklearn library function, we can easily realize it.

Fig. 4 This picture shows the fitness of GM method with training data and test data

The curve reflects the characteristics of the data, but it is not smooth.

B. Adaptive polynomial regression

Polynomial regression makes curves smoother and easier to implement. However, it is not easy to determine the power of polynomials, so the cross-validation method is introduced, which can realize the adaptive determination of the

number of arbitrary input data, so as to improve the efficiency.

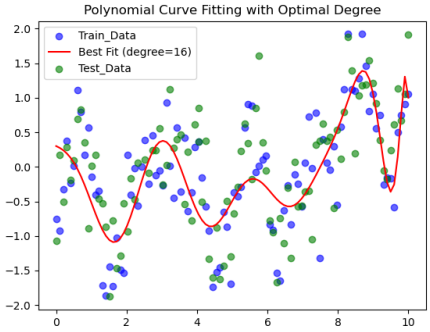

Fig. 5 This picture shows the fitness of GM method with training data and test data

C. Support Vector Regression

Support Vector machine Regression (SVR) is a regression algorithm based on support vector machines (SVMS) that is used to fit complex non-linear data while controlling the complexity of the model and reducing overfitting. sklearn library function provide a method to use it in coding. I use RBF kernel to deploy the model.

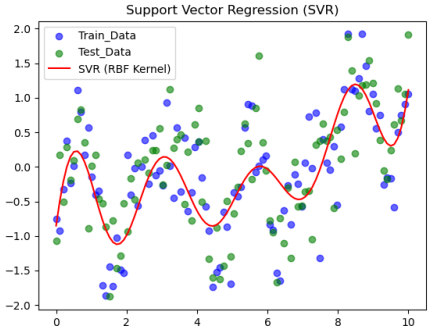

Fig. 6 This picture shows the fitness of GM method with training data and test data

III. Evaluations and Conclusion

Although there might be some points are off-line, it is clear that the 3 method filter some extreme noise and show the trait of data. And the result of SVR and APR seems to be much more smooth than RFR. Besides APR and SVR are almost the same in the middle part of the curve. It may because in the middle part more information are provided, satisfy the model to fit the curve.

In conclusion, SVR and APR are recommended to solving this problem.